# Optimized Search in Distributed Personal Content Systems

Niels Sluijs

Supervisor(s): Tim Wauters, Bart Dhoedt, Filip De Turck

## I. Introduction

Today's trend is to create and share *personal content*, such as text documents, music files, digital photos and digital movies. Different websites provide a framework to easily share personal content with each other, like Flickr[1] for digital photos and YouTube[2] for digital videos. The result of this trend is that each user's personal content archive grows explosively. Managing such a personal content archive has become a complex and time consuming task. This indicates the need of a personal content managing system that provides storage space *transparently*, with *small access time*, and available at *any place* and at *any time* to end-users.

Current systems (like YouTube and Flickr) that provide storage space for personal content are not capable to offer such a service in a *scalable*, *quality-aware* and in a *user-friendly* way. For instance file size is limited, restrictions are set on the file formats or do not provide the possibility to access personal content from different types of devices.

A networked solution that offers storage space to end-users in a transparent manner is a *Personal Content Storage Service* (PCSS), shown in Figure 1. For end-users the PCSS acts like a virtual hard disk, as if they access personal content on their local file system; with the advantage that the personal content is available at any time, from anywhere and from any device. As Figure 1 depicts, personal content can be accessed from e.g. a desktop computer, laptop, PDA (Personal Digital Agent), mobile phone, etc.

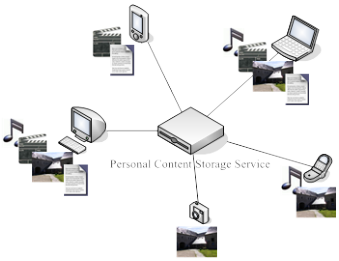


Figure 1: The Personal Content Storage Service will act as a virtual hard disk.

Besides offering transparent access to personal content, the PCSS will take care of storing content securely. This means that users are relieved from the task of making endlessly backups of their personal content.

In order to come to a successful deployment of a PCSS, research is needed for each different task of the PCSS. For instance, users want to search trough (their) personal content. Because of the relatively high number of users and content, performing search queries on a centralized architecture is not an option. To overcome this problem, *distributed* architectures have to be developed that offer a scalable solution. This article presents our research on the optimization of the response time of search queries on one hand and fast retrieval of the actual content on the other hand.

[1] http://www.flickr.com/
[2] http://www.youtube.com/

N. Sluijs is with the Department of Information Technology (INTEC), Ghent University (UGent), Ghent, Belgium. E-mail: Niels.Sluijs@UGent.be

## II. Search Queries

Due to the relatively large number of users and personal files, a distributed architecture is introduced for indexing and searching through personal content. Our overlay architecture allows indexing and searching through a large distributed dataset and is based on a *Distributed Hash Table* (DHT) [2]. Using a DHT, the index of a dataset is spread over nodes that participate in the DHT network; this decentralization improves the availability, flexibility and load balance. The average number of hops needed for a lookup in a DHT is $O(\log N)$, where $N$ is the number of nodes in the DHT. Reduction of the average number of hops implies a decrease in the average lookup latency. To improve the scalability we use a caching algorithm to decrease the average number of hops per lookup and evaluate it using the overlay simulator PlanetSim [1].

## III. Content Retrieval

Besides having an optimized distributed architecture to locate personal content, a user wants to be able to access the actual personal content quickly. The average latency decreases by using dimensioning algorithms [4] to optimize the capacity and location of storage servers, the amount of network bandwidth, etc. Additionally, we use a caching algorithm to further decrease the delay of retrieving the actual personal content [3].

## IV. Conclusion

In order to manage the explosive growth of personal content, a distributed network service is needed that offers access to personal content at any time, from anywhere and from any type of device. For successful deployment of a Personal Content Storage Service (PCSS), users need the ability to perform search queries. This article presents a distributed architecture that decreases the latency for search queries. On one hand we optimize the response time of obtaining the location where the personal content is stored and on the other hand we improve the delay of retrieving the actual content itself.

In the short term, we plan to focus on optimizing DHT solutions for multiple keywords searching. For the long term, we aim to study the usage of ontologies to enrich user queries.

## Acknowledgements

Our research is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Filip De Turck and Tim Wauters acknowledge the F.W.O.-V. (Fund for Scientific Research – Flanders) for their support through a postdoctoral fellowship.

## References

[1] J. P. Ahulló, P. G. López, M. S. Artigas, M. A. Arias, G. P. Aixalà, and M. Bruchmann, "PlanetSim: An extensible framework for overlay network and services simulations," Architecture and Telematic Services Research Group, Universitat Rovira i Vigili, Tarragona, Spain, DEIM-RR-08-002, 2008.

[2] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the evolution of peer-to-peer systems", in *Proceedings of the Twenty-First ACM Symposium on Principles of Distributed Computing*, Monterey, California, USA: 2002, pp. 233-242.

[3] N. Sluijs, K. Vlaeminck, T. Wauters, B. Dhoedt, F. De Turck, and P. Demeester, "Caching strategies for personal content storage grids", in *the 2007 International Conference on Parallel and Distributed Processing Techniques and Applications*, Volume 1, Las Vegas, Nevada, USA: 2007, pp. 396-404.

[4] K. Vlaeminck, T. Wauters, F. De Turck, B. Dhoedt, and P. Demeester, "Towards Transparent Personal Content Storage in Multi-service Access Networks", *Lecture Notes in Computer Science, Proceedings of EUC2007, the International Conference on Embedded and Ubiquitous Computing*, Volume: 4808, 2007, pp. 479-492.